the interior anchor-node parent cases to siblings and use the sibling anchor nodes as delimiters to generalize between. This is the case of anchor node parent with sibling delimiters.

Untangling is a problem that sometimes results when generalizing tabled structures. In step 126, the method may determine if there are any tangled nodes and then untangle any tangled nodes in step 128. The untangler method is described in more detail with respect to Figure 7. The problem manifests itself by generalized paths matching more than one item from each anchor node. In generalization, the idea is to create a general expression for a set of anchor nodes, but from each anchor node have very specific paths to the interior content pieces. So, if any interior path matches more than one item, there is a structural inconsistency and the structure that the user specified will not be represented in the RML output. These nodes need to be untangled by first counting the number of items each interior path matches as described below.

Once the generalization has proceeded through all the above steps, the replacements (or set of replacements for the case of untangling) are returned into the tree, replacing the generalized tag. The XSL writer then handles these as normal, without caring what the XPath expressions contain or whether they've been generalized to. Now, the path combining method in accordance with the invention will be described in more detail.

Figure 6 illustrates more details of the path combiner step 120 of the method shown in Figure 5. In particular, in step 130, the path combiner may match up the node path in each example. There are four cases, based on whether either of the following two statements are true: 1) the paths have been generalized before; and 2) the HTML is inconsistent as will be described. In step 132, the method determines if there are more paths. If there are no more paths, then the

method may compute the replacement element in the general paths and the path combiner method has been completed. If there are more paths, the method may determine if the paths have been generalized before in step 136. If the paths have been generalized before, it becomes more difficult to do that and instead the previously computed predicates are compared and concatenated with an 'or' operator to generalize the paths in step 140. If the paths have not been generalized before, it is a simple matter to attempt to take the HTML into consideration and try to find common attributes of nodes present in the paths in step 142.

In step 142, the method determines if the HTML is consistent. If the HTML is not consistent, the paths can be generalized on a step-by-step basis, considering each of the path elements independent of the rest in step 144. Otherwise, a method may figure out to what extent they are consistent and use set logic to figure out what is common between the paths for the remaining inconsistent part in step 146. That part of the algorithm relies upon an XSLT extension. In step 148, the generalized path are retrieved and the method is completed. Now, the node untangler method in accordance with the invention will be described in more detail.

Figure 7 illustrates more details of the node untangler step 128 of the method shown in Figure 5. In particular, as described above, the untangling problem manifests itself by generalized paths matching more than one item from each anchor node. Thus, these nodes need to be untangled by first counting the number of items each interior path matches. For a tangled node, the interior nodes will all match the same number of items. These paths are re-generalized by recovering the original paths to the examples, enumerating them by 1) the anchor node they are relative to, 2) the location of the path in the example structure, and 3) the item number. Then,

for each coordinate of (path,item) the paths are generalized across all anchor nodes. Then a number of replacements are used instead of just one, and this number is equal to the number of item numbers the tangled paths pointed to. In more detail, in step 150, the method may find all anchor nodes in the XHTML. In step 152, the method determines if there are any more anchor nodes. If there are more anchor nodes, then the method discovers the number of elements in each of the path matches in step 154 and indexes the paths by the location, anchor number and element number is step 156. The method then loops back to step 152 to determine if there are more anchor nodes.

If there are no more anchor nodes, then the method may combine the paths with the same element numbers in step 158 and create a predetermined number, N, of replacement elements in step 160. Thus, the untangling process in accordance with the invention is completed. Now, several examples of atomics or groups of atomics that may be generalized in accordance with the invention will be described to help better understand the invention.

In operation, a user may select an atomic or groups of atomics as examples of the groups or atomics that should be generalized. Based on the examples provided by a user and how these examples organize sections of XHTML, there are several very useful cases of how the generalizer should proceed in computing the mapping. These include:

Case 1. Generalizing atomics within a group (with a single group)

Case 2. Generalizing atomics within a group (with multiple groups)